plots as given in Figures 19 and in 20 can be used to identify the best fragments of threading models. Indeed, there are very strong correlations between the lowest mobility and the best structural fidelity (to the target structure) of the model chain fragments. This may have some other applications, where assessment of the reliability of various parts of a model structure is needed.

Summary and Conclusion

In this example, the invention again was shown to be useful in predicting medium- to low-resolution protein structures based on homology or sequence-structure compatibility. Here, the initial alignment between the target and template was generated by a threading procedure. Of course, alignments also can be obtained by other means, *e.g.*, from sequence alignments. Such templates are used to guide Monte Carlo simulations that employ a reduced protein chain representation built using pseudoatoms to represent the side chain center of mass of the various amino acid residues of a protein or protein domain. In contrast to the method of example 1, the pseudoatoms of the SICHO model used here took also took account of alpha-carbon atoms, in addition to the corresponding side chains. This alternate embodiment of the model proved capable of making large structural rearrangements that, in about a third of studied cases, lead to qualitative improvements in the initial poor models. In some other cases, despite a huge decrease in the RMSD between the model and the target native structure, the final model was still not satisfactory. The analysis of the simulation trajectories allows for the plausible identification of those cases where the final model improves qualitatively with respect to the initial, threading-based model.

The present invention is useful for large-scale protein structure and function prediction. Using the invention, it is possible to identify the biochemical function of a protein function having a model with a 5-6 Å backbone RMSD.[7,8] Certainly, it would be much more difficult, if not impossible, to make such an identification for a model with an 8 Å Cα RMSD from native polypeptide. For example, the model of

plastocyanin (2pcy) generated above had its four copper-binding residues much closer to their native position than predicted by the threading-based model. Thus, having a structural template of this active site (*e.g.*, an FSD), the model structure can be identified with high fidelity as a copper-binding protein. The results above show that for many new or known proteins (*e.g.*, those identified in the course of high throughput nucleic acid sequencing programs), the invention can be used to identify their function(s). The invention also complements sequence-based and threading methods, and provides a basis for improving initially poor and incomplete models. Additionally, the invention is also complementary to standard homology modeling tools, enabling homology modeling in those cases where the template is structurally very far from the target structure.

References (Example 2 only)

1. Altschul, S. F., Madden, T. L., Schaefer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res.* **25**, 3389-3402.

2. Aszodi, A. & Tylor, W. R. (1996). Homology modeling by distance geometry. *Folding & Design* **1**, 325-34.

3. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer Jr., E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Simanouchi, T. & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J Mol. Biol.* **112**, 535-542.

4. Binder, K. (1991). The Monte Carlo Method in Condensed Matter Physics, Institut Für Physik, Johannes Gutenberg-Universität, Mainz.

5. Bowie, J. U., Luethy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three dimensional structure. *Science* **253**, 164-170.

6. Bryant, S. H. & Lawrence, C. E. (1993). An empirical energy function for threading protein sequence through folding motif. *Proteins* **16**, 92-112.

7. Fetrow, J., Godzik, A. & Skolnick, J. (1998). Functional analysis of the *Escherichia coli* genome using the sequence-structure-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J Mol. Biol.*, **282**, 703-711.

8. Fetrow, J. S. & Skolnick, J. (1998). Method for prediction of protein function from sequence using the sequence to structure to function paradigm with application to glutaredoxins/thioredoxins and T_1 ribonucleases. *J. Mol. Biol.*, **281**, 949-968.

9. Godzik, A., Skolnick, J. & Kolinski, A. (1992). A topology fingerprint approach to the inverse folding problem. *J. Mol. Biol.*, **227**, 227-238.

10. Godzik, A., Skolnick, J. & Kolinski, A. (1993). Regularities in interaction patterns of globular proteins. *Protein Eng.* **6**, 801-810.

11. Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Nat'l Acad. Sci. USA* **89**, 10915-10919.

12. Hobohom, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Protein Sci.* **1**, 409-417.

13. Hu, W.-P., Godzik, A. & Skolnick, J. (1997). On the origin of sequence-structure specificity. How does an inverse folding approach work? *Prot. Engng.* **10**, 317-331.

14. Jaroszewski, L., Pawlowski, K. & Godzik, A. (1998a). Multiple model approach: Exploring the limits of comparative modeling. *J. Molecular Modeling*.

15. Jaroszewski, L., Rychlewski, L., Zhang, B. & Godzik, A. (1998b). Fold prediction by a hierarchy of sequence, threading, and modeling methods. *Protein Sci.* **7**, 1431-1440.

16. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature* **358**, 86-89.

17. Kolinski, A., Jaroszewski, L., Rotkiewicz, P. & Skolnick, J. (1998). An efficient Monte Carlo model of protein chains. Modeling the short-range correlations between side groups centers of mass. *J. Phys. Chem.* **102**, 4628-4637.